PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G06F 15/16, 17/00, 17/30

A1

(11) International Publication Number: WO 00/20982

(43) International Publication Date: 13 April 2000 (13.04.00)

(21) International Application Number: PCT/US99/22966

(22) International Filing Date: 1 October 1999 (01.10.99)

(30) Priority Data:
60/102,831 2 October 1998 (02.10.98) US

(71) Applicant *(for all designated States except US)*: NCR CORPORATION [US/US]; 101 W. Schantz Avenue, Dayton, OH 45479 (US).

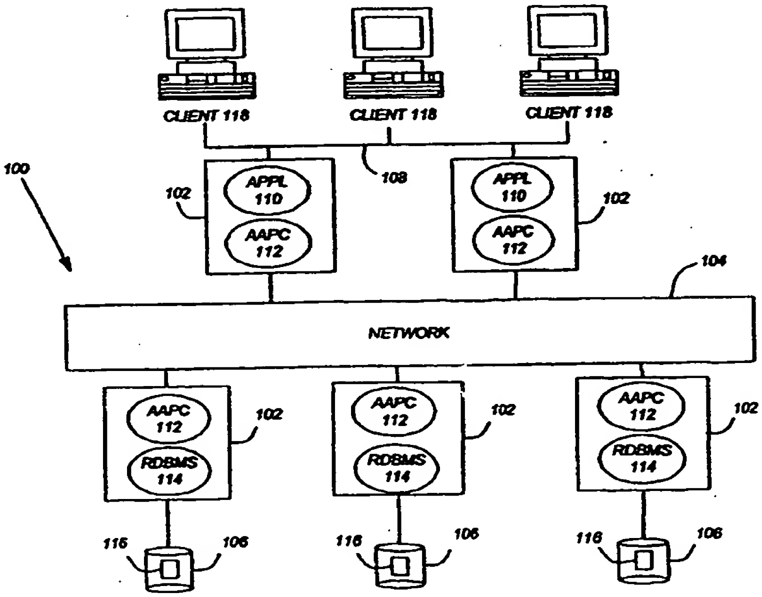
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: MILLER, Timothy, Edward [US/US]; 32668 Hupa Drive, Temecula, CA 92592 (US). TATE, Brian, Don [US/US]; 314 Skyridge Lane, Escondido, CA 92026 (US). HILDRETH, James, Dean [US/US]; 1545 Chandelle Lane, Fallbrook, CA 92028 (US). BRYE, Todd, Michael [US/US]; 12387 Briardale Way, San Diego, CA 92128 (US). ROLLINS, Anthony, Lowell [US/US]; 12502 Pacato Circle South, San Diego, CA 92128 (US). PRICER, James, Edward [US/US]; 2614 Winningham Road, Chapel Hill, NC 27516 (US). ANAND, Tej [US/US]; 71 Pond View Lane, Chappaqua, NY 10514 (US).

(74) Agents: STOVER, James, M.; NCR Corporation, 101 W. Schantz Avenue, Dayton, OH 45479 (US) et al.

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: SQL-BASED ANALYTIC ALGORITHMS



(57) Abstract

A method, apparatus, and article of manufacture for performing data mining applications in a relational database management system. At least one analytic algorithm (110) is performed by a computer directly against a relational database (116), wherein the analytic algorithm includes SQL statements performed by the relational database management system (114) and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relational database.

SQL-BASED ANALYTIC ALGORITHMS

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims the benefit under 35 U.S.C. Section 119(e) of the co-pending and commonly-assigned U.S. provisional patent application Serial No. 60/102,831, filed October 2, 1998, by Timothy E. Miller, Brian D. Tate, James D. Hildreth, Miriam H. Herman, Todd M. Brye, and James E. Pricer, entitled Teradata Scalable Discovery, which application is incorporated by reference herein.

This application is also related to the following co-pending and commonly-assigned utility patent applications:

Application Serial No. - -/ - - -, - - -, filed on same date herewith, by Brian D. Tate, James E. Pricer, Tej Anand, and Randy G. Kerber, entitled SQL-Based Analytic Algorithm for Association, attorney's docket number 8219,

Application Serial No. - -/ - - -, - - -, filed on same date herewith, by James D. Hildreth, entitled SQL-Based Analytic Algorithm for Clustering, attorney's docket number 8220,

Application Serial No. - -/ - - -, - - -, filed on same date herewith, by Todd M. Brye, entitled SQL-Based Analytic Algorithm for Rule Induction, attorney's docket number 8221,

Application Serial No. - -/ - - -, - - -, filed on same date herewith, by Brian D. Tate, entitled SQL-Based Automated Histogram Bin Data Derivation Assist, attorney's docket number 8222,

Application Serial No. - -/ - - -, - - -, filed on same date herewith, by Brian D. Tate, entitled SQL-Based Automated, Adaptive, Histogram Bin Data Description Assist, attorney's docket number 8223,

Application Serial No. PCT/US99/ - - - - -, filed on same date herewith, by Timothy E. Miller, Brian D. Tate, Miriam H. Herman, Todd M. Brye, and Anthony L. Rollins, entitled Data Mining Assists in a Relational Database Management System, attorney's docket number 8224,

Application Serial No. - -/ - - -, - - -, filed on same date herewith, by Todd M. Brye, Brian D. Tate, and Anthony L. Rollins, entitled SQL-Based Data Reduction Techniques for Delivering Data to Analytic Tools, attorney's docket number 8225, question typically involves probing the database through an iterative series of ad hoc or multidimensional queries until the root cause of the condition is discovered. Examples include Sales Analysis, Inventory Analysis or Production Analysis.

Stage three is the predicting stage, which tries to determine what will happen. As stage two users become more sophisticated, they begin to extend their analysis to include prediction of unknown events. For example, "Which end-users are likely to buy a particular product", or "Who is at risk of leaving for the competition?" It is difficult for humans to see or interpret subtle relationships in data, hence as data warehouse users evolve to sophisticated predictive analysis they soon reach the limits of traditional query and reporting tools. Data mining helps end-users break through these limitations by leveraging intelligent software tools to shift some of the analysis burden from the human to the machine, enabling the discovery of relationships that were previously unknown.

Many data mining technologies are available, from single algorithm solutions to complete tool suites. Most of these technologies, however, are used in a desktop environment where little data is captured and maintained. Therefore, most data mining tools are used to analyze small data samples, which were gathered from various sources into proprietary data structures or flat files. On the other hand, organizations are beginning to amass very large databases and end-users are asking more complex questions requiring access to these large databases.

Unfortunately, most data mining technologies cannot be used with large volumes of data. Further, most analytical techniques used in data mining are algorithmic-based rather than data-driven, and as such, there are currently little synergy between data mining and data warehouses. Moreover, from a usability perspective, traditional data mining techniques are too complex for use by database administrators and application programmers, and are too difficult to change for a different industry or a different customer.

Thus, there is a need in the art for data mining applications that directly operate against data warehouses, and that allow non-statisticians to benefit from advanced mathematical techniques available in a relational environment.

SUMMARY OF THE INVENTION

To overcome the limitations in the prior art described above, and to overcome other limitations that will become apparent upon reading and understanding the present specification, the present invention discloses a method, advanced analytic processing capabilities for data mining applications are placed where they belong, i.e., close to the data. Moreover, the results of these analytic processing capabilities can be made to persist within the database or can be exported from the database. These analytic processing capabilities and their results are exposed externally to the RDBMS by an application programmable interface (API).

According to the preferred embodiment, the data mining process is an iterative approach referred to as a "Knowledge Discovery Analytic Process" (KDAP). There are six major tasks within the KDAP:

1. Understanding the business objective.
2. Understanding the source data available.
3. Selecting the data set and "pre-processing" the data.
4. Designing the analytic model.
5. Creating and testing the models.
6. Deploying the analytic models.

The present invention provides various components for addressing these tasks:

- An RDBMS that executes Structured Query Language (SQL) statements against a relational database.
- An analytic Application Programming Interface (API) that creates scalable data mining functions comprised of complex SQL statements.
- Application programs that instantiate and parameterize the analytic API.
- Analytic algorithms utilizing:
  - Extended ANSI SQL statements,
  - a Call Level Interface (CLI) comprised of SQL staterments and programmatic iteration, and
  - a Data Reduction Utility Program comprised of SQL statements and programmatic iteration.
- An analytical logical data model (LDM) that stores results from and information about the advanced analytic processing in the RDBMS.
- A parallel deployer that controls parallel execution of the results of the analytic algorithms that are stored in the analytic logical data model.

node 102, causes the node 102 to perform the steps necessary to execute the steps or elements of the present invention.

Those skilled in the art will recognize that the exemplary environment illustrated in FIG. 1 is not intended to limit the present invention. Indeed, those skilled in the art will recognize that other alternative hardware environments may be used without departing from the scope of the present invention. In addition, it should be understood that the present invention may also apply to other computer programs than those disclosed herein.

LOGICAL ARCHITECTURE

FIG. 2 is a block diagram that illustrates an exemplary logical architecture of the AAPC 112, and its interaction with the APPL 110, RDBMS 114, relational database 116, and Client 118, according to the preferred embodiment of the present invention. In the preferred embodiment, the AAPC 112 includes the following components:

- An Analytic Logical Data Model (LDM) 200 that stores results from the advanced analytic processing in the RDBMS 114,
- One or more Scalable Data Mining Functions 202 that comprise complex, optimized SQL statements that perform advanced analytic processing in the RDBMS 114,
- An Analytic Application Programming Interface (API) 204 that provides a mechanism for an APPL 110 or other component to invoke the Scalable Data Mining Functions 202,
- One or more Analytic Algorithms 206 that can operate as standalone applications or can be invoked by another component, wherein the Analytic Algorithms 206 comprise:
    - Extended ANSI SQL 208 that can be used to implement a certain class of Analytic Algorithms 206,
    - A Call Level Interface (CLI) 210 that can be used when a combination of SQL and programmatic iteration is required to implement a certain class of Analytic Algorithms 206, and
    - A Data Reduction Utility Program 212 that can be used to implement a certain class of Analytic Algorithms 206 where data is first reduced using SQL followed by programmatic iteration.

In still another example, a Client 118 interacts with the APPL 110, which invokes one or more Analytic Algorithms 206 either directly or via the Analytic Algorithm API 214. The results would be stored as an analytic model within an Analytic LDM 200 in the RDBMS 114.

The overall goal is to significantly improve the performance, efficiency, and scalability of data mining operations by performing compute and/or I/O intensive operations in the various components. The preferred embodiment achieves this not only through the parallelism provided by the MPP computer system 100, but also from reducing the amount of data that flows between the APPL 110, AAPC 112, RDBMS 114, Client 118, and other components.

Those skilled in the art will recognize that the exemplary configurations illustrated and discussed in conjunction with FIG. 2 are not intended to limit the present invention. Indeed, those skilled in the art will recognize that other alternative configurations may be used without departing from the scope of the present invention. In addition, it should be understood that the present invention may also apply to other components than those disclosed herein.

Scalable Data Mining Functions

The Scalable Data Mining Functions 202 comprise complex, optimized SQL statements that are created, in the preferred embodiment, by parameterizing and instantiating the corresponding Analytic APIs 204. The Scalable Data Mining Functions 202 perform much of the advanced analytic processing for data mining applications, when performed by the RDBMS 114, without having to move data from the relational database 116.

The Scalable Data Mining Functions 202 can be categorized by the following functions:

- Data Description: The ability to understand and describe the available data using statistical techniques. For example, the generation of descriptive statistics, frequencies and/or histogram bins.
- Data Derivation: The ability to generate new variables (transformations) based upon existing detailed data when designing an analytic model. For example, the generation of predictive variables such as bitmaps, ranges, codes and mathematical functions.

The Analytic Algorithms 206 provide data analysts with an unprecedented option to train and apply "machine learning" analytics against massive amounts of data in the relational database 116. Prior techniques have failed as their sequential design is not optimal in an RDBMS 114 environment. Because the Analytic Algorithms 206 are implemented in Extended ANSI SQL 208, through the CLI 210, and/or by means of the Data Reduction Utility Program 212, they can therefore leverage the scalability available on the MPP computer system 100. In addition, taking a data-driven approach to analysis, through the use of complete Extended ANSI SQL 208, allows people other than highly educated statisticians to leverage the advanced analytic techniques offered by the Analytic Algorithms 206.

Extended ANSI SQL

As mentioned above, Analytic Algorithms 206 that are completely data driven, such as affinity analysis, can be implemented solely in Extended ANSI SQL 208. Typically, these type of algorithms operate against a set of tables in the relational database 116 that are populated with transaction-level data, the source of which could be point-of-sale devices, automated teller machines, call centers, the Internet, etc. The SQL statements used to process this data typically build relationships between and among data elements in the tables. For example, the SQL statements used to process data from point-of-sale devices may build relationships between and among products and pairs of products. Additionally, the dimension of time can be added in such a way that these relationships can be analyzed to determine how they change over time. As the implementation is solely in SQL statements, the design takes advantage of the hardware and software environment of the preferred embodiment by decomposing the SQL statements into a plurality of sort and merge steps that can be executed concurrently in parallel by the MPP computer system 100.

Call-Level Interface

As mentioned above, Analytic Algorithms 206 that require a mix of programmatic iteration along with Extended ANSI SQL statements, such as inductive inference, can be implemented using the CLI 210. Whereas the SQL approach is appropriate for business problems that are descriptive in nature, inference problems are predictive in nature and typically require a training phase where the APPL 110 "learns" various rules based upon the data description, define the characteristics of data stored in the relational database 116, as well as metadata that determines how the RDBMS 114 performs the advanced analytic processing. The Analytic LDM 200 also stores processing results from this advanced analytic processing, which includes both result tables and derived data for the Scalable Data Mining Functions 202, Analytic Algorithms 206, and the Parallel Deployer 216. The Analytic LDM 200 is a dynamic model, since the logical entities and attributes definitions change depending upon parameterization of the advanced analytic processing, and since the Analytic LDM 200 is updated with the results of the advanced analytic processing.

Logic of the Preferred Embodiment

Flowcharts which illustrate the logic of the preferred embodiment of the present invention are provided in FIGS. 3, 4 and 5. Those skilled in the art will recognize that this logic is provided for illustrative purposes only and that different logic may be used to accomplish the same results.

Referring to FIG. 3, this flowchart illustrates the logic of the Scalable Data Mining Functions 202 according to the preferred embodiment of the present invention.

Block 300 represents the one or more of the Scalable Data Mining Functions 202 being created via the API 204. This may entail, for example, the instantiation of an object providing the desired function.

Block 302 represents certain parameters being passed to the API 204, in order to control the operation of the Scalable Data Mining Functions 202.

Block 304 represents the metadata in the Analytic LDM 200 being accessed, if necessary for the operation of the Scalable Data Mining Function 202.

Block 306 represents the API 204 generating a Scalable Data Mining Function 204 in the form of a data mining query based on the passed parameters and optional metadata.

Block 308 represents the Scalable Data Mining Function 204 being passed to the RDBMS 114 for execution.

Referring to FIG. 4, this flowchart illustrates the logic of the Analytic Algorithms 206 according to the preferred embodiment of the present invention.

Block 400 represents the Analytic Algorithms 206 being invoked, either directly or via the Analytic Algorithm API 214.

computer, such as a mainframe, minicomputer, or personal computer, could be used to implement the present invention.

In summary, the present invention discloses a method, apparatus, and article of manufacture for performing data mining applications in a relational database management system. At least one analytic algorithm is performed by a computer directly against a relational database, wherein the analytic algorithm includes SQL statements performed by the relational database management system and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relational database.

The foregoing description of the preferred embodiment of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. It is intended that the scope of the invention be limited not by this detailed description, but rather by the claims appended hereto.

8. The computer-implemented system of claim 1, wherein the analytic algorithm is implemented by a Data ReductionUtility Program that reduces data from the relational database in bulk using SQL followed by a non-SQL iterative program..

9. The computer-implemented system of claim 8, wherein the Data Reduction Utility Program provides a sequence of Extended ANSI SQL followed by programmatic iteration.

10. A method for performing data mining applications, comprising:

(a) managing a relational database stored on one or more data storage devices connected to a computer; and

(b) performing at least one analytic algorithm in the computer, wherein the analytic algorithm includes SQL statements performed by a relational database management system directly against the relational database and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relational database.

11. An article of manufacture comprising logic embodying a method for performing data mining applications, comprising:

(a) managing a relational database stored on one or more data storage devices connected to a computer; and

(b) performing at least one analytic algorithm in the computer, wherein the analytic algorithm includes SQL statements performed by a relational database management system directly against the relational database and optional programmatic iteration, and the analytic algorithm creates at least one analytic model within an analytic logical data model from data residing in the relational database.

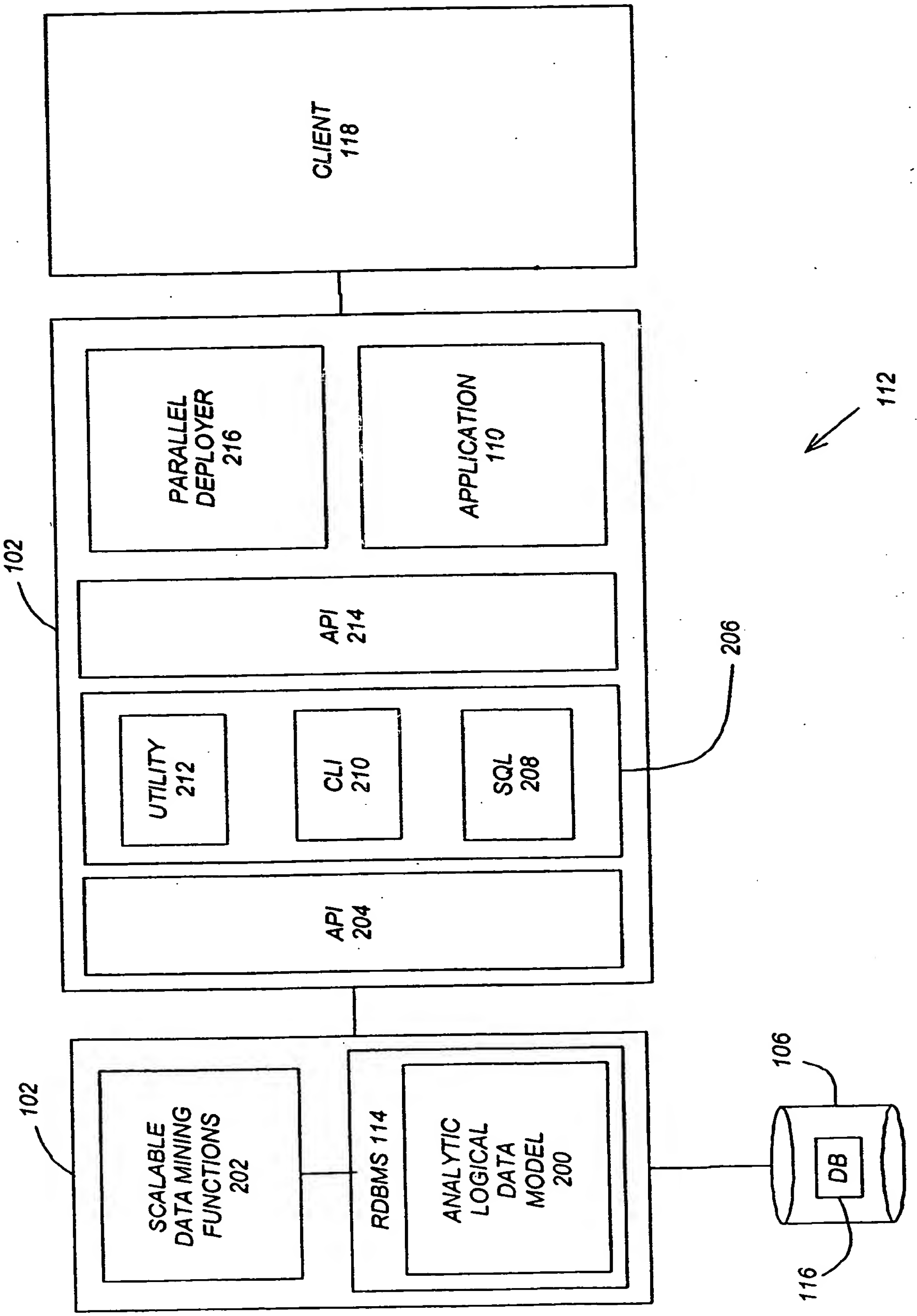

CLIENT 118
PARALLEL DEPLOYER 216
APPLICATION 110
API 214
UTILITY 212
CLI 210
SQL 208
API 204
102
206
112
SCALABLE DATA MINING FUNCTIONS 202
RDBMS 114
ANALYTIC LOGICAL DATA MODEL 200
102
DB
106
116

FIG. 2

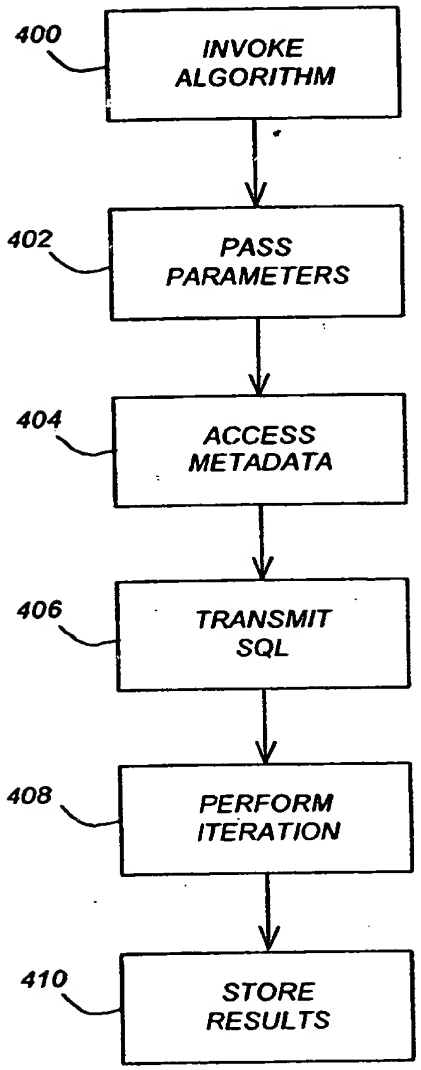

400
INVOKE ALGORITHM
402
PASS PARAMETERS
404
ACCESS METADATA
406
TRANSMIT SQL
408
PERFORM ITERATION
410
STORE RESULTS

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/22966

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G06F 15/16, 17/00, 17/30
US CL :707/2, 4, 100, 102
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 707/2, 4, 100, 102

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WEST, EAST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5, 412,806 A (DU et al.) 02 May 1995, the entire paper is relevant	1-11
Y	US 5,590,322 A (HARDING et al.) 31 December 1996, the entire paper is relevant	1-11
Y	US 5,799,310 A (ANDERSON et al.) 25 August 1998, the entire paper is relevant	1-11
Y	US 5, 806,066 A (GOLSHANI et al.) 08 September 1998, the entire paper is relevant	1-11

□ Further documents are listed in the continuation of Box C. □ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

08 DECEMBER 1999

Date of mailing of the international search report

23 DEC 1999

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
THUY PARDO James R. Matthews
Telephone No. (703) 305-1091